Scientometrics (2014) 101:1233–1252
DOI 10.1007/s11192-014-1279-6

Citation impact prediction for scientific papers using stepwise regression analysis

Tian Yu · Guang Yu · Peng-Yu Li · Liang Wang

Received: 17 November 2013 / Published online: 30 March 2014

Abstract Researchers typically pay greater attention to scientific papers published within the last 2 years, and especially papers that may have great citation impact in the future. However, the accuracy of current citation impact prediction methods is still not satisfactory. This paper argues that objective features of scientific papers can make citation impact prediction relatively accurate. The external features of a paper, features of authors, features of the journal of publication, and features of citations are all considered in constructing a paper's feature space. The stepwise multiple regression analysis is used to select appropriate features from the space and to build a regression model for explaining the relationship between citation impact and the chosen features. The validity of this model is also experimentally verified in the subject area of Information Science & Library Science. The results show that the regression model is effective within this subject.

Keywords Scientific paper · Citation impact prediction · Feature space · Multiple regression

Introduction

The scientific paper is the basic unit of scientometric research. As carriers of knowledge, published papers influence scientific communication and progress. The citation impact is

T. Yu (✉) · G. Yu · L. Wang
School of Management, Harbin Institute of Technology, Harbin, People's Republic of China
e-mail: yutian.hit@gmail.com

G. Yu
e-mail: yug@hit.edu.cn

L. Wang
e-mail: hitwangliang@hit.edu.cn

P.-Y. Li
School of Education, Harbin Institute of Technology, Harbin, People's Republic of China
e-mail: lipengyu.hit@gmail.com

defined as the sum of citations that reference a given paper, and is a widely used measure of scientific impact for publications. Individual papers, journals, scientists, institutions, etc. have been evaluated or even ranked based on citation impact (Hargens and Schuman 1990).

In the current era of knowledge explosion, researchers can obtain a large volume of papers in a given research subject conveniently. A count of papers listed on Web of Science reveals that a researcher in the subject area of Information Science & Library Science reading two papers daily would take at least 100 years to finish all of the articles. However, as the reading time of an individual researcher is limited, a researcher does not want to waste time reading papers of no significance. For papers that have already been published for more than 5 years, we can easily evaluate each paper's citation impact by its citation count. However, for papers that only have been published one or 2 years, it is difficult to predict their future citation impact. As papers published within a short period of time normally cover the current research hotspots and trends, researchers pay more attention to these publications to raise the novelty of their studies. It is thus important to predict the citation impact of papers published within the last 2 years.

The objective of this study is to understand, for the benefit of academic researchers engaged in evaluating scientific papers, the interactions between features of scientific papers and their citation impact. The future citation impact of scientific papers could be predicted accurately through an evaluation of their features. Previous studies have shown that accepted high-quality papers have a strong capacity for knowledge diffusion in the first 5 years after publication (Glänzel et al. 2003; Aksnes 2003), which implies that the citation impact of a paper 5 years after publication is an important manifestation of the paper's quality (Wang et al. 2011, 2012). Therefore, in our research, we predict paper citation impact after 5 years of publication to determine paper quality.

In this paper, we analyze relevant features of scientific papers to find the relationship between these features and citation impact, and to thereby predict the number of citations after the first 5 years of publication. Section 2 discusses some related work. Section 3 presents the method for description and establishes the feature space used to characterize scientific papers. Section 4 explains the data and methodology used in this research. In Sect. 5, we apply correlation analysis and regression analysis to find a model that explains the relationship between the features and citation impact. Section 6 provides a conclusion.

Related work

The present study shows that paper citation impact may be influenced by four factors: authors, published journals, research fields and the quality of the papers themselves.

Author characteristics

As scientific papers are produced by researchers involved in scientific exploration, the characteristics of authors are indirectly reflected in the papers. The Matthew effect suggests that authors prefer to cite not only the papers that help their studies but also those that are written by reputable scholars. There is some evidence to support the notion that author reputation affects citation counts. Stewart (1983) found the author reputation effect on citation count to be present in geophysical research. Danell (2011) concluded that, to a certain degree, the impact of scientific work can be predicted using information on author past performance. Moreover, some interesting author characteristics such as gender and institutional affiliation are assumed to be determinants in citation allocation (Van Dalen and Henkens 2001; Prpić 2002; Peñas and Willett 2006).

Journal characteristics

It is generally accepted that journals (and their editors) with good reputations attract high-quality papers. This implies that potentially high-quality papers are submitted to core journals, while lower-quality papers will be submitted to second-tier journals. Because a prestigious journal signals to researchers that the papers that it publishes are of high quality, researchers read the papers published in core journals first, while papers published in second-tier journals tend to be ignored. Van Dalen and Henkens (1999, 2001) provided some evidence that articles published in core journals received considerably more citations than articles published in second-tier journals, and the majority of articles in second-tier journals remained uncited in the 5 years following publication. Researchers have applied a zero-inflated negative binomial regression model on papers published in nanoscience and nanotechnology journals in the Web of Science from 2007 to 2009 and found the publishing journal impact factor to be the most reliable determinant of citation counts (Didegah and Thelwall 2013).

Field characteristics

Garfield (1979) stressed that the research field must be taken into account when comparing between citation counts generated in different research fields because the 'citation potential' can vary significantly from one field to another. Some researchers highlighted that delineations between research fields could be artificially and ambiguously defined (Boyack and Klavans 2011). Researchers have attempted to achieve field normalization with different methods to solve the problem of the delineation of appropriate sets of scientific papers for the comparison. Radicchi et al. (2008) and Radicchi and Castellano (2012) considered a new relative indicator $c_f = c/c_0$, where c is the number of citations to an article and c_0 is the average number of citations per article for the discipline. Their study provided the validation of c_f as an effective indicator for citation performance across disciplines. Moed (2010) explored a new indicator of journal citation impact—source normalized impact per paper (SNIP)—to allow for direct comparisons of sources in different subject fields. Furthermore, some researchers have begun to consider using non-parametric statistics rather than central tendency statistics, as the latter does not provide accurate representations of citation distributions because the distributions are skewed. Leydesdorff and Bornmann (2011) applied the newly defined 'Integrated Impact Indicator' (I3) as a measure of citation, and Leydesdorff (2012) elaborated on statistics such as the top-10 % of the most-highly cited papers on the basis of the fractional counting of citations, which may provide an alternative to the current IF.

Article characteristics

It has been well documented that the quality of a scientific paper is one of the most important factors influencing citation impact. Van Dalen and Henkens (2005) stated that paper quality could be approximated by the impact and speed with which knowledge is disseminated in the scientific community. Citations indicate impact in the literature. Although citations are not entirely accurate measures of impact, they are nevertheless indicators that bring some insight to scientific communication. The speed with which a paper is disseminated in the scientific community is measured by the timing of the first citation. Researchers have also explored whether external factors of scientific papers such as language, length and type may affect citation counts (Portes 1998; Van Dalen and Henkens 2001).

As predicting citations is extremely difficult and complicated (Feitelson and Yovel 2004), most researchers focus on citation analysis and citation behavior rather than citation count prediction (Fu and Aliferis 2010). Even so, some statistical models have been established to predict the future citation behavior of publications based on the above features. Glänzel and Schubert (1995) presented a non-homogeneous birth-process model. Burrell (2001, 2003) presented a theory for a stochastic model on the citation process in the presence of obsolescence to predict the future citation patterns of individual papers. Using machine learning methods, Fu and Aliferis (2010) and Fu et al. (2013) presented models that accurately predict the citation counts of biomedical publications with a mixture of content-based and bibliometric features. Recently, Wang et al. (2011, 2012) established a high-cited papers' prediction model using a machine learning tool. Comparing the number of citations with the mean citation rate of the particular research field, the authors identified high-, medium- and low-cited papers and established a prediction model of highly cited papers using a pattern recognition technique. They also found that paper citation features after the first 5 years of publication were of great importance in identifying highly cited papers. However, the above researchers all used classification methods to predict whether a paper may be highly cited in future. Classification output is discrete and boundaries between classes tend to be ambiguous. In Wang et al.'s study papers were divided into three classes by citation count. However, there should be a minor difference between a highly cited paper with proportionally fewer citations and a medium-cited paper with a large relative citation count. Moreover, their study used the citation features of a paper after the first 5 years of publication to predict its class after 10 years. Currently, the update rate of scientific papers is high, e.g. the cited half-life of journals in subject of Information Science & Library Science is about six. With the development of databases such as Google Scholar and Web of Science, researchers can easily access recently published papers and prefer to read and cite the papers published within the last 5 years to ensure novelty in their work. It seems that the total number of citations to one paper 10 years after published is significantly less important. Therefore, we attempt to predict the citation impact of papers after the first 5 years of publication by regression analysis, which is introduced to perform more a detailed classification prediction of citation count in our study.

Feature space of scientific papers

The scientific paper can be described as a vector collection of multi-dimensional information including references, author(s), a research field, etc. In other words, there are multi-dimensional features of papers. The feature space X of scientific papers can be defined below:

$$X = \{x_1, x_2, x_3, \ldots, x_n\}$$

where x_i ($i = 1, 2, \ldots, n$) is the feature of papers. The citation impact y of a scientific paper is defined as the total number of citations.

The features describing a scientific paper are divided into four types: the external features of a paper, features of authors, features of the journal of publication and the features of citations. External features of a paper include the article type, language, publication date, number of references, etc. To facilitate convenient comparison, we only select papers of the article type and which were published in 2007. According to the Matthew effect, author reputation and journal reputation are likely to influence the total number of citations. Several scientometric indicators, such as the total number of publications and citations and citations per journal paper, may characterize the scientific publications quantitatively. A series of journal evaluation indicators from JCR and Eigenfactor^{TM} metrics are used to characterize the respective quality and impact of journals and their editorial boards. Moreover, features of citations such as the first-cited age and capacity of knowledge diffusion during the first 2 years after publication are used to measure the quality of scientific papers.

Table 1 lists the features of scientific papers. These features are simple indicators that are both widely accepted and easily accessible. The reciprocal of the first-cited age is used as the first-cited age in this study. The reason is for this is that some papers have never been cited in Web of Science. To facilitate comparison, the first-cited age for these papers could be defined as positive infinity. The reciprocal of the first-cited age could be in the range 0–1. The paper with high value of the reciprocal of the first-cited age should diffuse more rapidly in the scientific community.

Methods

Data preparation

Our study is based on data provided by Thomson ISI. The 2007 version of the JCR has indexed 56 journals in the subject of Information Science & Library Science. Due to time constraints on data collection, we selected the first 20 journals whose indicators were completed in JCR according to the JCR list of 2007 (listed in Table 2). Using these ISI products, we collected data until January of 2012.

As the papers selected were all published in 2007, the features of the journals in which the papers were published were obtained directly from the 2007 version of the JCR.

The online version of Web of Science provides an analysis tool called ‘Analyze Results’ for analyzing paper characteristics. Using this tool, we first extracted the citations published in the first 2 years after publication from all citations, and then we could be able to analyze the features of citations in the period of the first 2 years after publication.

Features of authors can be identified in several steps. First, to obtain an author’s total number of publications, we consult ‘Distinct Author Record Sets’, which is a discovery tool in Web of Science showing sets of papers likely written by the same person. Second, we exclude the papers written by other authors with the same name from all publications. This process is highly labor intensive as it requires separating the papers published by different authors with the same name and then extracting the papers published by the desired author based on the author’s affiliation, address, email and so on. Third, we exclude the papers published from 2007 to 2012. The month of publication is ignored here for the convenience of statistical analysis. Finally, we calculate the features of each author before publication of the selected paper. In addition, we use the country of the first corresponding author as the location of the author’s institution, which could be statistically analysed only because it is a text feature.

Table 1 Features of scientific papers

Feature type	Feature description	Label
External features of a paper	The type (the document type for each selected paper is the article)	
	The number of references	x_1
Features of authors	The number of authors	x_2
	The country of the author's institution (text type features)	
	The *h* index of the first author before publication of this paper	x_3
	The number of papers published by the first author before this paper	x_4
	The total citations to the papers published by the first author before this paper	x_5
	The average citations to the papers published by the first author before this paper	x_6
	The maximum *h* index of the authors before publication of this paper	x_7
	The maximum number of papers published by the authors before this paper	x_8
	The maximum total citations to the papers published by the authors before this paper	x_9
	The maximum average citations to the papers published by the authors before this paper	x_{10}
Features of citations	The reciprocal of the first-cited age of this paper	x_{11}
	The total citations to this paper in its first 2 years after publication	x_{12}
	The number of countries citing this paper in its first 2 years after publication	x_{13}
	The number of paper types citing this paper in its first 2 years after publication	x_{14}
	The number of journals citing this paper in its first 2 years after publication	x_{15}
	The number of subjects citing this paper in its first 2 years after publication	x_{16}
Features of the journal of publication	The total citations to the journal	x_{17}
	The impact factor of the journal	x_{18}
	The 5-year impact factor of the journal	x_{19}
	The immediacy index of the journal	x_{20}
	The number of papers published in the journal this year	x_{21}
	The cited half-life of the journal	x_{22}
	The Eigen factor score of the journal	x_{23}
	The article influence score of the journal	x_{24}

We have identified the features of 1,025 papers published in 20 journals of Information Science & Library Science using the ISI database. Data collection was completed in January 2012.

Analysis method

As the focus of this paper is on the relationship between citation impact and the features shown in Table 1, we adopt multiple regression analysis to obtain the scoring function for the feature set.

Table 2 List of 20 journals studied from Information Science & Library Science

No.	Abbreviated journal title	ISSN
1	Aslib Proc	0001–253X
2	Coll Res Libr	0010–0870
3	Gov Inform Q	0740–624X
4	Inform Manage-Amster	0378–7206
5	Inform Process Manag	0306–4573
6	Inform Res	1368–1613
7	Inform Soc	0197–2243
8	Inform Syst J	1350–1917
9	Inform Syst Res	1047–7047
10	Int J Geogr Inf Sci	1365–8816
11	Int J Inform Manage	0268–4012
12	J Acad Libr	0099–1333
13	J Am Med Inform Assn	1067–5027
14	J Am Soc Inf Sci Tec	1532–2882
15	J Doc	0022–0418
16	J Health Commun	1081–0730
17	J Inf Sci	0165–5515
18	J Inf Technol	0268–3962
19	J Libr Inf Sci	0961–0006
20	J Manage Inform Syst	0742–1222

Multiple regression analysis is an important branch of applied statistics. Gibbons (1982) first used the multivariate regression model method to measure the effect of new information on asset prices. The method not only extracts the important information hidden in massive data sets but can also use variables to predict and control a specific variable (Kleinbaum et al. 1998). In regression analysis, an output variable is called the dependent variable, and the variables that influence the dependent variable are called independent variables. The dependent variable changes in response to changes in the independent variables. Hence, in our regression analysis the number of citations is considered the dependent variable and these 24 features shown in Table 1 are the independent variables. MATLAB 2009 for Windows was used to complete the majority of our calculations.

To account for the possibility of collinearity in some of the features, we first performed a correlation analysis of article features and citation impact, which yielded the optimal regression equation through multiple linear stepwise regression analysis. Stepwise regression is a technique for choosing variables to include in a multiple regression model. This not only guarantees the validity and importance of the chosen variables but also reduces additional error introduced by the redundant variables. The main approaches are forward selection, backward elimination and bidirectional elimination. Forward selection, also called forward stepwise regression, starts with an equation with no variables. At each step the technique involves adding the variable with the highest F statistic or lowest p value until there are none left. Backward elimination, also called backward stepwise regression, begins with all of the variables in the equation and removes the least significant variables until all remaining variables are statistically significant. Bidirectional elimination involves a combination of the above, testing for variables to be included or excluded at each step. In stepwise regression, the variables ending up in the final equation signify the best combination of independent variables to predict the dependent variable. In this research, bidirectional elimination is applied to obtain the optimal model.

To ensure model stability, we ran an additional ten-fold cross-validation process and obtained an optimal model from ten regression models via stepwise regression analysis (Kohavi 1995). There are several kernel steps to this process:

(1) The sample set D is divided into ten subsets (d_1, d_2, …, d_{10}), where $D = d_1 \cup d_2 \cup \cdots \cup d_{10}$ and $d_i \cap d_j = \varphi(i \neq j, i = 1, 2, \ldots, 10$ and $j = 1, 2, \ldots, 10)$,
(2) $q = 1$,
(3) If $q > 10$, execute (7),
(4) The subset d_q is defined as the test set, and the rest of the subsets are defined as the training set,
(5) Stepwise regression is executed for the training set and the regression function_q is obtained,
(6) $q = q + 1$, execute (3),
(7) Select the model with the smallest MAPE and SDEW as the optimal model.

Results and discussion

Correlation analysis of features

Our data set contains 1,025 papers published across 20 journals in 2007, and the accumulated total number of citations to these papers is 7,232. Figure 1 shows that citations are skewed in the papers to total number of citations y distribution, suggesting that most papers are cited only a few times. As this finding conforms to overall trends in the subject area of Information Science & Library Science, the data we selected are valid.

The summary of descriptive statistics for all 1,025 papers is provided in this section. Overall, most of the papers possess 10–50 references. More than 80 % of all papers are written by one to three authors. Of the authors of 452 papers, nearly 45 % of the papers come from American institutions. Authors from England, Korea, Spain and Canada also published a large number of papers in 2007.

The features x_3, x_4, x_5 and x_6 reflect the reputation of first author. In our data set, approximately 75 % of the first author h index values before publication of paper (x_3) are not more than two; roughly 70 % of the number of papers published by the first author before paper (x_4) are not more than four; about 80 % of the total citations to the papers published by the first author before paper (x_5) are lower than 60; about 70 % of the average citations to each paper published by the first author before paper (x_6) are lower than nine. This implies that about half of all researchers in the field are new and of low prestige. Moreover, the features x_7, x_8, x_9 and x_{10} reflect the highest co-author reputations. For our data set, about 80 % of the maximum h index values of the authors before publication of paper (x_7) are not more than seven; about 60 % of the maximum number of papers published by the authors before paper (x_8) are not more than seven; over 80 % of the maximum total citations to the papers published by the authors before paper (x_9) are lower than 1,250; about 80 % of the maximum average citations to each paper published by the authors before paper (x_{10}) are lower than 30.

In our data more than 50 % of all papers were first cited within their first 2 years of publication, and approximately 75 % were first cited in the first 3 years. Moreover, highly

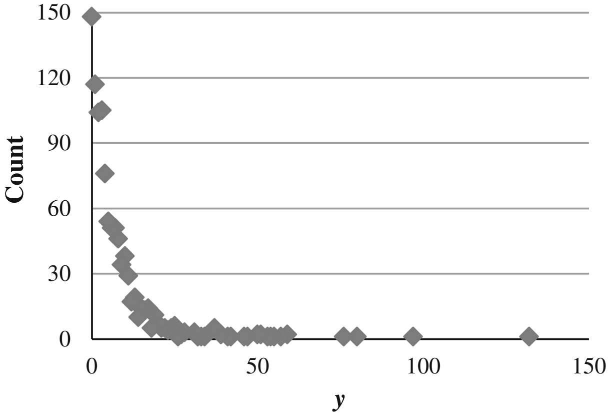

Fig. 1 Distribution of 1,025 articles across the total number of citations *y*

cited papers are strongly correlated with citations in the first year after publication, and the corresponding paper impact and speed of knowledge diffusion are high.

To capture the features of scientific papers that affect citation impact, correlation analysis is used in this section. To suit the characteristics and distribution of our data, we used the Spearman correlation coefficient to measure the relationship between citation impact *y* and the 24 features. The Spearman correlation coefficient is a non-parametric measure of statistical dependence between two variables. The measure assesses how well the relationship between two variables can be described using a monotonic function.

External features of a paper

Table 3 illustrates the Spearman correlation coefficient between citation impact and external features. Since the correlation coefficient indicates the non-directional relationship between two variables, the Spearman correlation matrix is a symmetric matrix. This result shows that the correlation coefficient between the number of references and citation impact is close to 0.5, which is a medium value. This result demonstrates that the number of references could influence the citation impact.

Features of authors

Spearman correlation coefficients between citation impact and features of authors are shown in Table 4. The result shows that a strong correlation is observed between any two of these features with the exception of the number of authors. The correlations between the number of authors and the other seven features are very low. In addition, these features could affect the citation impact slightly, and their influence on citation impact is lower than their influence on external features.

Features of citations

Table 5 shows the Spearman correlation coefficient between citation impact and features of citations. As the correlation coefficients are all beyond 0.65, these features are significantly associated with the citation impact.

Table 3 Correlation between citation impact and external features of a paper

	Correlation between vectors of values	
	y	x_1
y	1.000	0.406**
x_1	0.406**	1.000

This is a symmetric matrix

** Significant at the 0.01 level

Features of the journal of publication

The Spearman correlation coefficients between citation impact and the features of the journal of publication are shown in Table 6. Their correlation coefficients lie at roughly 0.3, which indicates that the citation impact is more related to features of the journal of publication than to the features of authors. The correlation coefficients between citation impact and the number of papers published in the journal is less than 0.2, meaning that it is less effective to improve a journal's reputation by increasing the number of papers published.

Stepwise multiple linear regression model on citation impact

In the above section, the effects of 24 features on citation impact were examined. However, these relationships are not precise and specific. Hence, in this section, we introduce regression techniques to obtain the exact relationships between citation impact and these features.

Logically, it is necessary to prove that the features of scientific papers do in fact influence the number of citations before examining the relationship between the features and citation impact. As we have shown in the above section that these features do influence citation impact, we can apply the regression analysis directly.

The distributions of these features for the 1,025 papers have been obtained. This indicates that we can use the linear regression model, the most common model in regression analysis, to explore the relationship between the features and citation impact.

To facilitate data analysis, we normalize all variables. Here y_i represents the citation impact of the paper i in the first 5 years after publication, and $x_{i,j}$ represents the value of the feature j of the paper i. The normalized dependent variable y_i* and independent variables $x_{i,j}$* are as follows:

$$y_i^* = \frac{y_i}{y_B}, \quad i = 1, 2, \ldots, 1025$$

$$x_{i,j}^* = \frac{x_{i,j}}{x_{jB}}, \quad i = 1, 2, \ldots 1025, \; j = 1, 2, \ldots, 24$$

where $y_B = \bar{y} = \frac{1}{1025}\sum_{i=1}^{1025} y_i$, $x_{jB} = \bar{x}_j = \frac{1}{1025}\sum_{i=1}^{1025} x_{i,j}$.

Multiple linear regression analysis is used to estimate the parameters of the linear function based on the given data. The regression model is determined from a set of known citation impacts of 1,025 papers, which provide the determined values of these 24 features. We use these normalized variables to conduct the linear regression analysis with the help of computer software (MATLAB).

Table 4 Correlation between citation impact and the features of authors

	Correlation between vectors of values									
	y	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
y	1.000	0.212**	0.175**	0.082**	0.218**	0.258**	0.287**	0.190**	0.321**	0.369**
x_2	0.212**	1.000	0.071	0.081*	0.096*	0.118**	0.369**	0.386**	0.406**	0.377**
x_3	0.175**	0.071	1.000	0.881**	0.960**	0.885**	0.675**	0.596**	0.633**	0.573**
x_4	0.082**	0.081*	0.881**	1.000	0.838**	0.703**	0.586**	0.654**	0.542**	0.413**
x_5	0.218**	0.096*	0.960**	0.838**	1.000	0.958**	0.665**	0.579**	0.672**	0.651**
x_6	0.258**	0.118**	0.885**	0.703**	0.958**	1.000	0.625**	0.490**	0.657**	0.710**
x_7	0.287**	0.369**	0.675**	0.586**	0.665**	0.625**	1.000	0.900**	0.949**	0.794**
x_8	0.190**	0.386**	0.596**	0.654**	0.579**	0.490**	0.900**	1.000	0.848**	0.603**
x_9	0.321**	0.406**	0.633**	0.542**	0.672**	0.657**	0.949**	0.848**	1.000	0.895**
x_{10}	0.369**	0.377**	0.573**	0.413**	0.651**	0.710**	0.794**	0.603**	0.895**	1.000

This is a symmetric matrix

* Significant at the 0.05 level; ** significant at the 0.01 level

Table 5 Correlation between citation impact and the features of citations

	Correlation between vectors of values						
	y	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
y	1.000	0.716**	0.714**	0.686**	0.662**	0.704**	0.690**
x_{11}	0.716**	1.000	0.853**	0.830**	0.841**	0.838**	0.823**
x_{12}	0.714**	0.853**	1.000	0.944**	0.928**	0.965**	0.927**
x_{13}	0.686**	0.830**	0.944**	1.000	0.941**	0.959**	0.934**
x_{14}	0.662**	0.841**	0.928**	0.941**	1.000	0.948**	0.923**
x_{15}	0.704**	0.838**	0.965**	0.959**	0.948**	1.000	0.957**
x_{16}	0.690**	0.823**	0.927**	0.934**	0.923**	0.957**	1.000

This is a symmetric matrix

** Significant at the 0.01 level

Table 6 Correlation between citation impact and the features of the journal of publication

	Correlation between vectors of values								
	y	x_{17}	x_{18}	x_{19}	x_{20}	x_{21}	x_{22}	x_{23}	x_{24}
y	1.000	0.353**	0.366**	0.366**	0.281**	0.184**	0.334**	0.318**	0.346**
x_{17}	0.353**	1.000	0.644**	0.807**	0.787**	0.747**	0.649**	0.847**	0.889**
x_{18}	0.366**	0.644**	1.000	0.887**	0.670**	0.337**	0.537**	0.499**	0.801**
x_{19}	0.366**	0.807**	0.887**	1.000	0.779**	0.416**	0.439**	0.799**	0.942**
x_{20}	0.281**	0.787**	0.670**	0.779**	1.000	0.684**	0.421**	0.639**	0.729**
x_{21}	0.184**	0.747**	0.337**	0.416**	0.684**	1.000	0.358**	0.571**	0.456**
x_{22}	0.334**	0.649**	0.537**	0.439**	0.421**	0.358**	1.000	0.296**	0.512**
x_{23}	0.318**	0.847**	0.499**	0.799**	0.639**	0.571**	0.296**	1.000	0.847**
x_{24}	0.346**	0.889**	0.801**	0.942**	0.729**	0.456**	0.512**	0.847**	1.000

This is a symmetric matrix

** Significant at the 0.01 level

The correlation coefficient matrix of the 24 features reveals high correlations between some features. Since introducing all of the variables into the regression model can cause the multi-collinearity problem, we applied a stepwise regression analysis for choosing good variables from all of variables in generating the predictor team. To obtain reliable and stable results, a ten-fold cross-validation was used, yielding ten regression models through stepwise regression analysis. The number of features involved in each model is shown in Table 7.

Table 7 shows the variables involved in each regression model. In the table, ‘1’ means that the model includes this variable, and ‘0’ means that the model does not include this variable. We examined the prediction accuracy of the ten regression models separately. The *F* values and RMSE of the ten models show that ten models provide a better fit to the real data. Furthermore, MAPE and SDEW show the prediction accuracy of each model. The model with the smallest MAPE and SDEW was selected as the optimal model; this was described in Table 8.

Table 8 shows that several of the chosen independent variables involved in the optimal model are significant at the 0.01 level. The *R*, *R*-squared and adjusted *R*-squared for this model are 0.822, 0.676 and 0.674, respectively. This shows that the regression model can explain the relationships between the features and citation impact. The results of an additional ANOVA show that the model is statistically significant. Table 8 also reveals that all variance inflation factors (VIF) for the selected features fall below 1.5. There is virtually no collinearity in this model.

Therefore, based on the results of our regression, the regression equation for our analysis can be written as:

$$\hat{y}^* = 0.30x_1^* + 0.10x_2^* + 0.03x_5^* - 0.15x_{11}^* + 0.70x_{12}^* + 0.24x_{19}^* - 0.13$$

where $\hat{y}^*$ represents the estimated or predicted value of the normalized citation impact y*. The six coefficients in front of the normalized features are regression coefficients, which denote how much change occurs to the normalized citation impact y* when one normalized feature changes by one unit, while the others remain constant. Moreover, our regression equation includes a constant term.

These results can be interpreted to demonstrate that x_1, x_2, x_5, x_{12} and x_{19} have positive impacts on research performance (citation impact) and x_{11} has a negative impact on the research performance (citation impact). In this regression model, external feature of the paper, features of authors, features of the journal of publication and features of citations all significantly contribute to citation impact prediction.

The prediction results of our regression equation of the test sample are shown in Figs. 2 and 3.

After establishing the functional relationship between the features and citation impact, we used test data to examine the regression model's validity. The number of citations accumulated after the test papers' first 5 years of publication was determined to compare the predictive values with the actual citation impact values. Although there are some errors, Figs. 2 and 3 show that the prediction results of our regression model are good for most of test samples, proving that the regression model is relatively effective. Hence, paper citation impact after 5 years of publication can be predicted using objectively assessed factors.

As each of the four types of features of scientific papers can affect the citation impact, our results demonstrate validity. Similar to the findings of previous studies (Van Dalen and Henkens 2001; Danell 2011), our results also indicate that author reputation and journal's rank can affect a scientific paper's citation impact. This provides the necessary basis enabling us to perform regression analyses. Moreover, we find that the number of references shows some relationship to paper citation impact. These findings were obtained through a regression model treating citation impact as a dependent variable, and the result has proven statistically significant.

Explanation for the regression model

In the above section, we obtained the optimal regression model describing the citation impact of one scientific paper. We found six features that play a significant role in affecting paper citation impact: the number of references (x_1), the number of authors (x_2), the total number of citations to the papers published by the first author before this paper (x_5), the reciprocal of the first-cited age of this paper (x_{11}), the total citations to this paper after the first 2 years of publication (x_{12}) and the 5-year impact factor of the journal (x_{19}). Our findings suggest that all four types of features describing scientific papers are significantly correlated with citation impact. The features do however vary by strength of correlation: the features of citations have the strongest influence, followed by the external features of a

Table 7 Variables involved in each regression model

No.	x_1*	x_2*	x_5*	x_9*	x_{10}*	x_{11}*	x_{12}*	x_{15}*	x_{18}*	x_{19}*	*F*	RMSE	MAPE	SDEW
1	1	0	0	1	0	0	1	0	0	1	324.84	0.84	0.71	0.78
2	1	1	1	0	0	1	1	0	0	1	288.60	0.82	**0.65**	**0.45**
3	1	0	0	0	1	1	1	1	0	1	244.90	0.80	0.71	0.49
4	1	0	1	0	1	0	1	0	0	1	242.70	0.77	0.65	0.51
5	1	0	1	0	0	1	1	0	0	1	321.96	0.84	0.66	0.56
6	1	0	0	0	1	0	1	0	1	0	246.87	0.82	0.68	0.54
7	1	1	0	0	0	1	1	0	0	1	302.38	0.82	0.73	0.68
8	1	0	0	0	1	1	1	1	0	1	302.10	0.83	0.81	0.64
9	1	0	0	0	0	0	1	0	0	1	328.89	0.81	0.84	0.70
10	1	0	0	1	1	1	1	0	0	1	336.00	0.83	0.81	0.67

Bold values indicate that the model with the smallest MAPE and SDEW is selected as the optimal model

Table 8 Regression coefficients of the model

Feature	*B*	Sig.	VIF
x_1*	0.30	0.00**	1.081
x_2*	0.10	0.00**	1.074
x_5*	0.03	0.00**	1.059
x_{11}*	−0.15	0.00**	1.121
x_{12}*	0.70	0.00**	1.464
x_{19}*	0.24	0.00**	1.436

** Significant at the 0.01 level

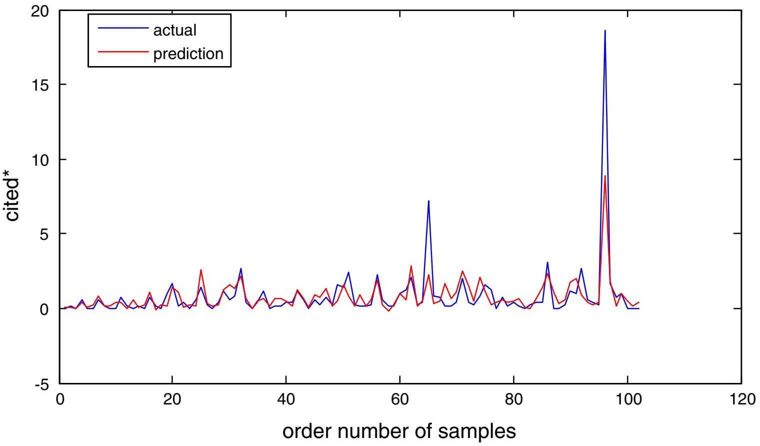

Fig. 2 Comparison between the regression results and the actual total citations of the test samples

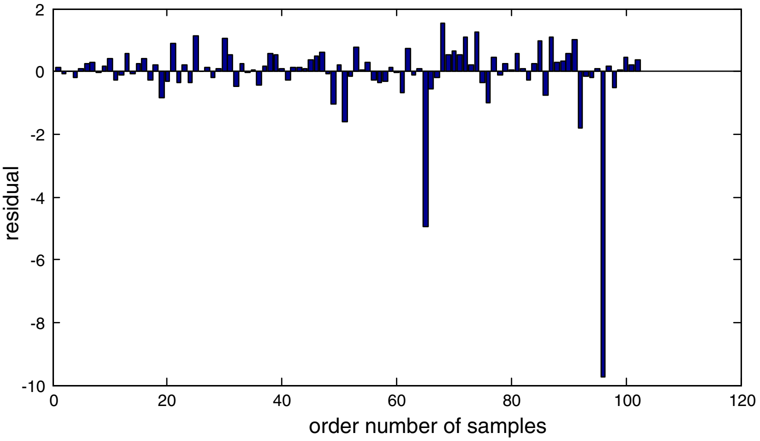

Fig. 3 Residual plots of the test samples

paper itself, features of the journal of publication and features of authors. Clearly, our research demonstrates that the features of citations determine the speed and breadth of knowledge diffusion in a citation network. One could approximate the contents and quality of a paper by these features (Van Dalen and Henkens 2005). It is expected that the features describing external appearance and internal quality of a paper make the greatest contributions to paper citation impact. In addition, the features of authors and of the journal of publication affect the citation impact to some extent. Compared with the features of authors, the features of the journal of publication have a more significant on improving citation impact.

A paper's quality

The regression model we obtained includes two features of citations. The first-cited age and the total citations to a paper after the first 2 years of publication measure the speed with which knowledge is diffused in scientific community and a degree of acceptance by peers and other professionals, respectively. Through correlation analysis, the correlation coefficients between citation impact and these two features both fall beyond 0.7. Our regression model finds these features to be the most important factors associated with citation impact.

According to Van Dalen and Henkens (2005), the quality of scientific papers is 'approximated by the impact and speed with which knowledge is disseminated in the scientific community'. Our research finds that the impact of a paper which knowledge is disseminated within the scientific community boils down to the total citations to the paper after the first 2 years of publication (x_{12}), the number of countries (x_{13}), types of papers (x_{14}), journals (x_{15}) and subjects (x_{16}) citing the paper after the first 2 years of publication and the speed with which an paper is disseminated in the scientific community is measured by the first-cited age (x_{11}). In the above section, our regression model includes the first-cited age and total citations to a paper after the first 2 years of publication, denoting that early citations are a good indication of quality and are closely related to paper citation impact.

Of the four types of features describing a scientific paper, external features of a paper, features of authors and features of the journal of publication can be measured immediately following paper publication. However, features of citations may only be measured after a paper has been published several years. To further determine the influence of these citation features on citation impact and predict citation impact for recently published scientific papers, we establish a citation impact model using all the features except for the citation features via stepwise multiple linear regression analysis. The R, R-squared and adjusted R-squared for the obtained regression model are 0.439, 0.193 and 0.177, respectively, indicating that the model cannot explain the relationship between citation impact and all of the features minus the features of citations. Although we attempted to predict the citation impact through other regression techniques, the models obtained failed to explain the relationship adequately. In short, we failed to determine an effective model for predicting citation impact using all features but features of citation. We confirm that the quality of a scientific paper is one of the most significant factors affecting citation impact (Van Dalen and Henkens 2005).

These two distinct regression results show that it is difficult to predict citation impact by these features for a recently published a paper. As a paper's age increases, more signs of citation impact become evident, making predicting its citation impact more feasible.

In fact, these two distinct regression results demonstrate the essential nature of citing behaviour. Citing behavior illustrates the continuity and inheritance during the development of science process. The probability of being cited depends on many factors, especially on the contents and quality of a paper. High quality papers usually receive more attention from research fellows, increasing the probability of being cited.

A paper's external features

The number of references in a paper, an external feature used to characterize scientific papers in our research, was found to have a significant influence on citation impact. This is likely a consequence of the author's knowledge of literatures in his field. The more literature a researcher reads, the more deeply he understands the current situation and trends of development of his research field. This is an effective strategy for enhancing the quality of one's research.

Journal and author reputation

The regression model we obtained includes features of authors and of the journal of publication, meaning that journal and author reputation are generally felt to play a role of some significance in attracting attention in science.

In terms of the features of authors, our optimal model includes two variables: the number of authors (x_2) and the total number of citations to the papers published by the first author before the selected paper (x_5). Echoing to the results of previous studies (Leimu and Koricheva 2005; Borsuk et al. 2009; Gazni and Didegah 2010), the number of authors could affect the paper's citation impact to a certain extent. Our results illustrate that in subject of Information Science & Library Science, papers published by many authors attract more attention. The results also explain authorship trends for journals in this subject area: the number of papers with single authorship is declining while the number of papers involving national and international collaboration is increasing (Sin 2011). Furthermore, the correlation analysis shows that the correlation coefficients between citation impact and author features fall at roughly 0.25. However, the influence of first-author features on citation impact is lower than that of highest prestige author features. This finding explains researchers' tendencies to cooperate with individuals of high prestige and impact. The regression model also shows that author reputation may in some cases exert only weak influence on citation impact. While papers may get noticed immediately based on author reputation, they may prove in the long run to be of little value.

With respect to the features of the journal of publication, our regression model suggests that journals are a major force in determining citation impact. Previous research has shown that papers published in core journals receive considerably more citations than papers published in second-tier journals (Van Dalen and Henkens 2001). A journal's reputation for originality and influence can attract high-quality papers. Potentially influential papers will be submitted first to journals that have a reputation for being influential. Hence, the scientific communication process reinforces a journal's reputation. Our regression coefficients show further that compared to author reputation, journal reputation has a greater influence on citation impact. To some extent, this finding emphasizes the considerable role of journal editors in shaping citation trends.

In addition, although journal and author reputation are correlated with citation impact, reading habits and the citation motivations of researchers are also significant factors affecting citation impact. Merton's Matthew effect (1968) is applicable here: researchers are more willing to read and cite the papers that are written by famous authors or published in core journals.

Overall, our results suggest that the characteristics of a scientific paper are very important factors in predicting its influence. In fact, citation impact is a complex phenomenon involving many explicit and implicit social and scholarly factors. Though the six variables included in our model may be the most apparent, there is a need to acknowledge the existence of other factors associated with citation impact.

Conclusions

In summary, our results suggest that a paper's citation impact can be predicted by objective scientometric indicators. External features of a paper, features of authors, features of the journal of publication, and features of citations are all involved in constructing a paper's feature space through the mathematical description method. Given that the information provided by these features may be redundant, the method of stepwise regression analysis was applied to select good variables from all of the features and build a model describing the relationship between the features and citation impact. Because citation potential can vary significantly across fields, papers were limited to those published in the subject of Information Science & Library Science to avoid error. With relative accuracy, we can predict paper citation impact after the first 5 years of publication in this subject.

Several important caveats should temper these conclusions. Most importantly, our research has shown the interesting relationship between various features and citation impact. By this we do not claim that these features cause citation impact, but rather that they are significant indicators. Although we believe the scientific paper to possess a multi-dimensional complex of features, in this study we only selected the features that were considered available and which could be obtained in a relatively simple and fast manner. We, for instance, did not consider the characteristics of the citing papers to be determinants of citation results. This assumption may cause the omission of some features. Limitations of the ISI database, our platform for data acquisition, such as data incompleteness are bound to be brought into this study. However, it is undeniable that the ISI is the largest comprehensive academic information resource database in the world that covers most subjects, and thus it is a good source of data for this research. Finally, the sample of scientific papers used in this analysis is quite limited and is strictly a reflection of papers published in the selected subject category. Our model therefore only applies to this subject category.

Despite these caveats, the findings of this study reveal interesting relationships between citation impact and the features of scientific papers. The feature space constructed by the selected features is effective in describing scientific papers. We must further consider the comprehensiveness and effectiveness of these features, which will involve many aspects of the open access status of the paper itself, the acceptable level of audiences, etc. Finally, the data must be larger and more comprehensive.

Acknowledgments This work was supported by the National Natural Science Foundation of China (Grant No. 70973031).

References

Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation, 12*(3), 159–170.

Borsuk, R. M., Budden, A. E., Leimu, R., Aarssen, L. W., & Lortie, C. J. (2009). The influence of author gender, national language and number of authors on citation rate in ecology. *Open Ecology Journal, 2*, 25–28.

Boyack, K. W., & Klavans, R. (2011). Multiple dimensions of journal specificity: Why journals can't be assigned to disciplines. In E. Noyons, P. Ngulube, & J. Leta (Eds.), *The 13th conference of the international society for scientometrics and informetrics* (Vol. I, pp. 123–133). Durban: ISSI, Leiden University and the University of Zululand.

Burrell, Q. L. (2001). Stochastic modelling of the first-citation distribution. *Scientometrics, 52*, 3–12.

Burrell, Q. L. (2003). Predicting future citation behavior. *Journal of the American Society for Information Science and Technology, 54*(5), 372–378.

Danell, R. (2011). Can the quality of scientific work be predicted using information on the author's track record? *Journal of the American Society for Information Science and Technology, 62*(1), 50–60.

Didegah, F., & Thelwall, M. (2013). Determinants of research citation impact in nanoscience and nanotechnology. *Journal of the American Society for Information Science and Technology, 64*(5), 1055–1064.

Feitelson, D., & Yovel, U. (2004). Predictive ranking of computer scientists using CiteSeer data. *Journal of Documentation, 60*(1), 44–61.

Fu, L. D., & Aliferis, C. F. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics, 85*(1), 257–270.

Fu, L. D., Aphinyanaphongs, Y., & Aliferis, C. F. (2013). Computer models for identifying instrumental citations in the biomedical literature. *Scientometrics, 97*(3), 871–882.

Garfield, E. (1979). *Citation indexing: Its theory and application in science, technology and humanities*. New York: Wiley.

Gazni, A., & Didegah, F. (2010). Investigating different types of research collaboration and citation impact: A case study of Harvard University's publications. *Scientometrics, 87*(2), 251–265.

Gibbons, M. R. (1982). Multivariate tests of financial models: A new approach. *Journal of Financial Economics, 10*(1), 3–27.

Glänzel, W., Schlemmer, B., & Thijs, B. (2003). Better later than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics, 58*(3), 571–586.

Glänzel, W., & Schubert, A. (1995). Predictive aspects of a stochastic model for citation processes. *Information Processing and Management, 31*(1), 69–80.

Hargens, L. L., & Schuman, H. (1990). Citation counts and social comparisons: Scientists' use and evaluation of citation index data. *Social Science Research, 19*(3), 205–221.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., & Nizam, A. (1998). *Applied regression analysis and other multivariable methods*. Pacific Grove: Brooks/Cole Publishing Company.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 2*(12), 1137–1143.

Leimu, R., & Koricheva, J. (2005). Does scientific collaboration increase the impact of ecological articles? *BioScience, 55*(5), 438–443.

Leydesdorff, L. (2012). Alternatives to the journal impact factor: I3 and the top-10 % (or top-25 %?) of the most-highly cited papers. *Scientometrics, 92*(2), 355–365.

Leydesdorff, L., & Bornmann, L. (2011). Integrated impact indicators (I3) compared with impact factors (IFs): An alternative design with policy implications. *Journal of the American Society for Information Science and Technology, 62*(7), 1370–1381.

Merton, R. K. (1968). The Matthew effect in science. *Science, 159*, 56–63.

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics, 4*(3), 265–277.

Peñas, C. S., & Willett, P. (2006). Brief communication: Gender differences in publication and citation counts in librarianship and information science research. *Journal of Information Science, 32*(5), 480–485.

Portes, A. (1998). Social capital: Its origins and applications in modern sociology. *Annual Review of Sociology, 24*, 1–24.

Prpić, K. (2002). Gender and productivity differentials in science. *Scientometrics, 55*(1), 27–58.

Radicchi, F., & Castellano, C. (2012). Testing the fairness of citation indicators for comparison across scientific domains: The case of fractional citation counts. *Journal of Informetrics, 6*(1), 121–130.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *PNAS, 105*(45), 17268–17272.

Sin, S. C. J. (2011). International coauthorship and citation impact: A bibliometric study of six LIS journals, 1980–2008. *Journal of the American Society for Information Science and Technology, 62*(9), 1770–1783.

Stewart, J. A. (1983). Achievement and ascriptive processes in the recognition of scientific articles. *Social Forces, 62*(1), 166–189.

Van Dalen, H. P., & Henkens, K. (1999). How influential are demography journals? *Population and Development Review, 25*(2), 229–251.

Van Dalen, H. P., & Henkens, K. (2001). What makes a scientific article influential? The case of demographers. *Scientometrics, 50*(3), 455–482.

Van Dalen, H. P., & Henkens, K. (2005). Signals in science-on the importance of signaling in gaining attention in science. *Scientometrics, 64*(2), 209–233.

Wang, M. Y., Yu, G., An, S., & Yu, D. R. (2012). Discovery of factors influencing citation impact based on a soft fuzzy rough set model. *Scientometrics, 93*(3), 635–644.

Wang, M. Y., Yu, G., & Yu, D. R. (2011). Mining typical features for highly cited papers. *Scientometrics, 87*(3), 695–706.